Methods for Mixtures

or the Subgroup Problem

S. Stanley Young
NISS

27Aug2012

Large Observational Data Sets

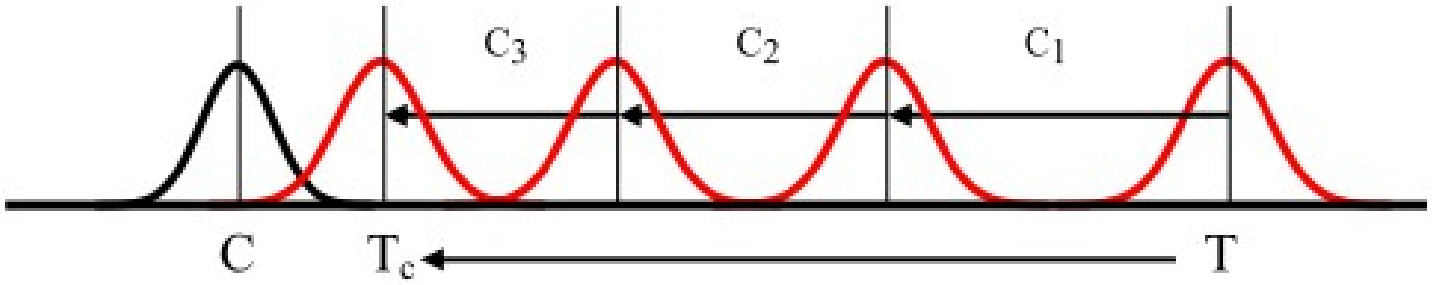
- Logistically awkward
- Prompt multiple questions
- Invite multiple modeling
- Subject to bias
- Subject to data staging variability

Discussion

- Obenchain
- Madigan
- Data – availability and quality
- To predict or to explain
- A blast from the past

The problem of big n

(a) Use confounding variables to reduce bias.



(b) As n get large the standard error of the mean gets small.

Obenchain

0. Design (see Rubin)
1. Cluster (number and method)
2. Local treatment differences within clusters
3. Distribution of LTDs vs simulation
4. LTD-Xvector, => Recursive partitioning

So simple, difficult to manipulate answers.

State of the Art, 1988

AMERICAN JOURNAL OF EPIDEMIOLOGY
Vol. 127, No. 3
Printed in U.S.A.

ASYMMETRIC STRATIFICATION

AN OUTLINE FOR AN EFFICIENT METHOD FOR CONTROLLING CONFOUNDING IN COHORT STUDIES

E. FRANCIS COOK AND LEE GOLDMAN

Madigan

1. Data staging flexibility
2. Total experimental variability
3. Repeatability issues
4. Average versus individual

Data

1. Much data is effectively private
2. Synthetic/simulated for methods evaluation
3. Public data sources (Heejung Bang)
4. OMOP
5. Sentinel – Congress/FDA
   (Fire, Ready, Aim)

Mixtures, Predict/Explain

Statistical Science
2010, Vol. 25, No. 3, 289–310
DOI: 10.1214/10-STS330

To Explain or to Predict?

Galit Shmueli

RP single tree versus forest

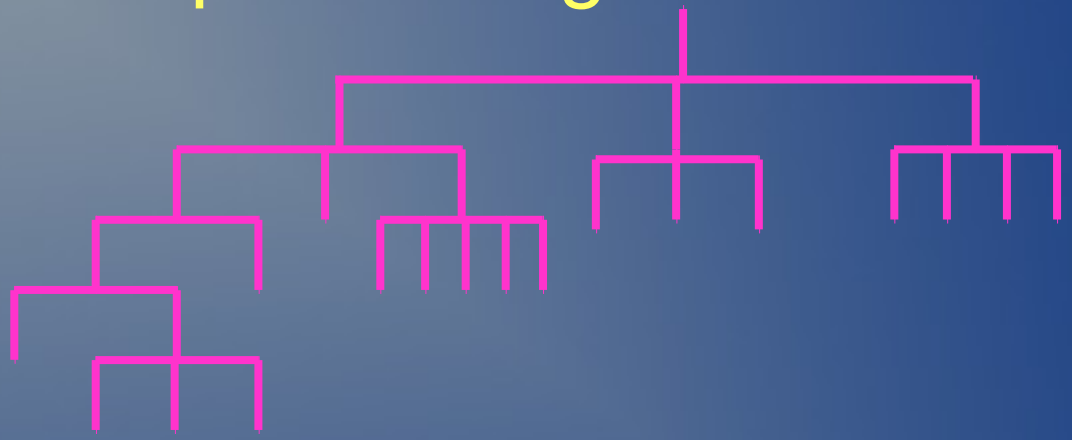
PCA versus NMF

MLR versus SVM

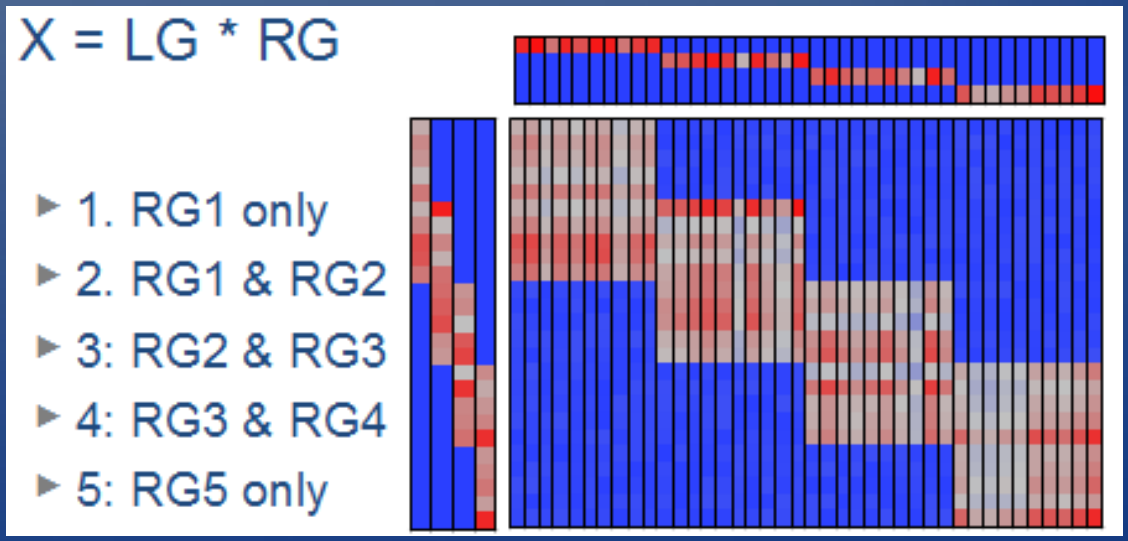
Etc.

Two methods for mixtures

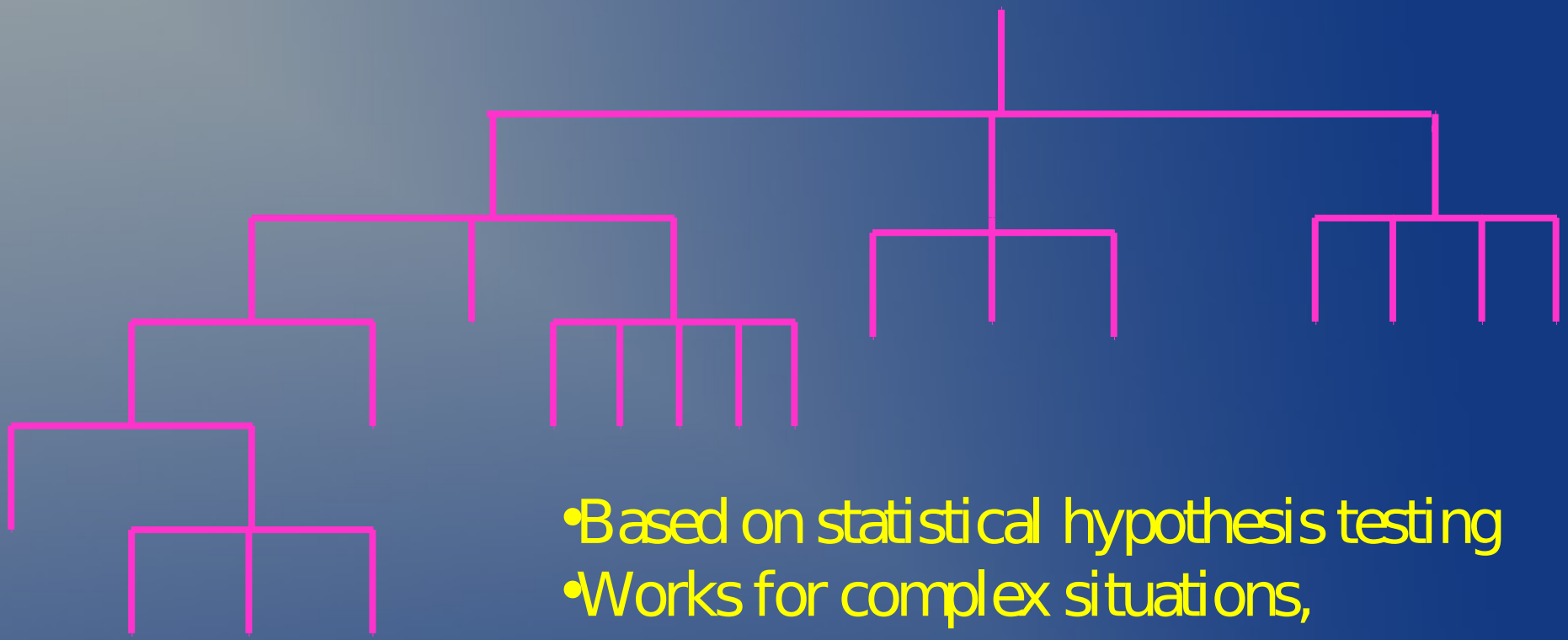
- Recursive partitioning



- Non-negative matrix factorization



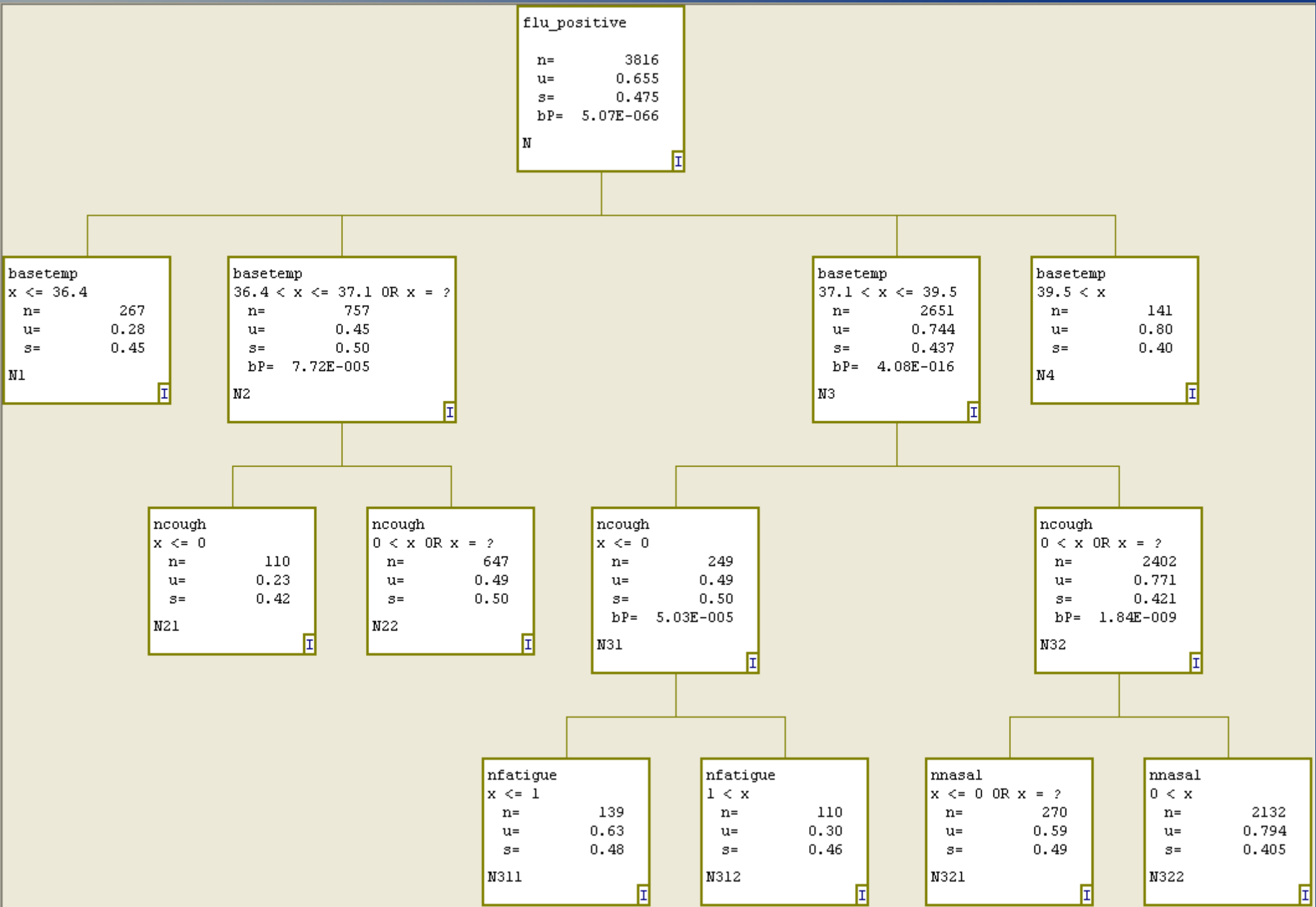
Recursive Partitioning – Finding Groups



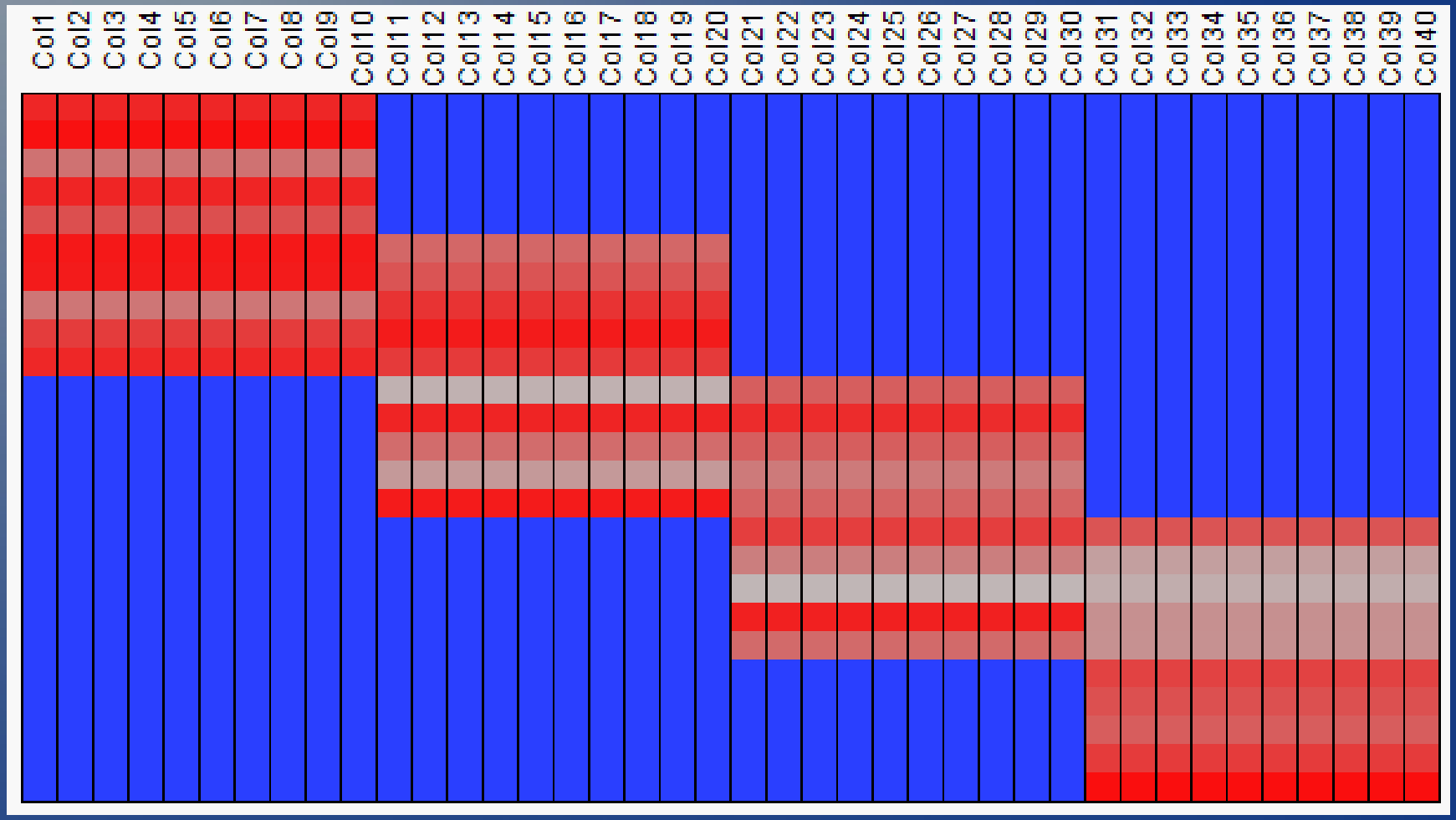
- Based on statistical hypothesis testing
- Works for complex situations, mixtures and interactions
- Statistical method easy to understand
- Excellent for subgroup analysis
- Handles more predictors than observations

Hawkins algorithms, Golden Helix, State of the Art RP

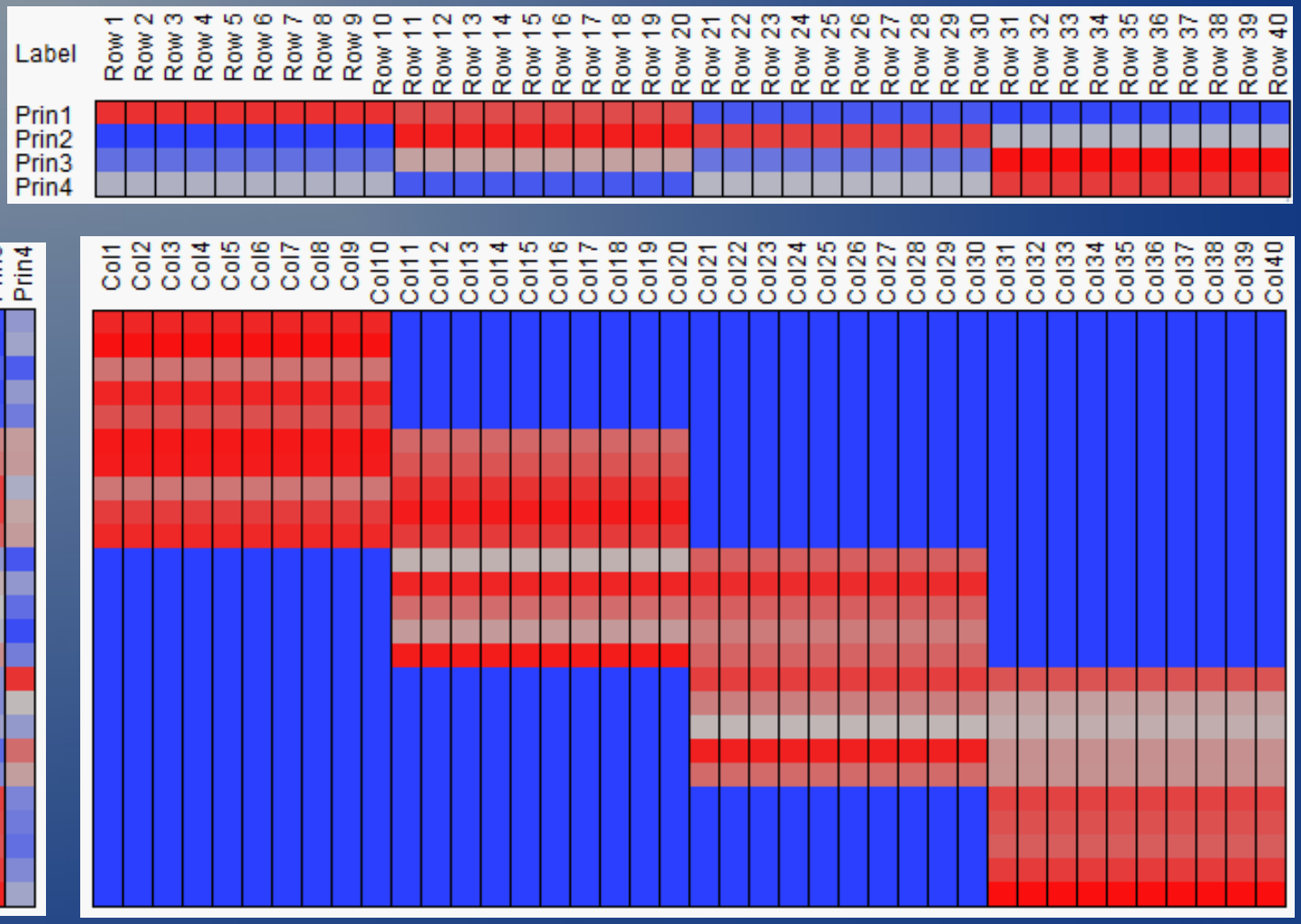
Partionator Tree



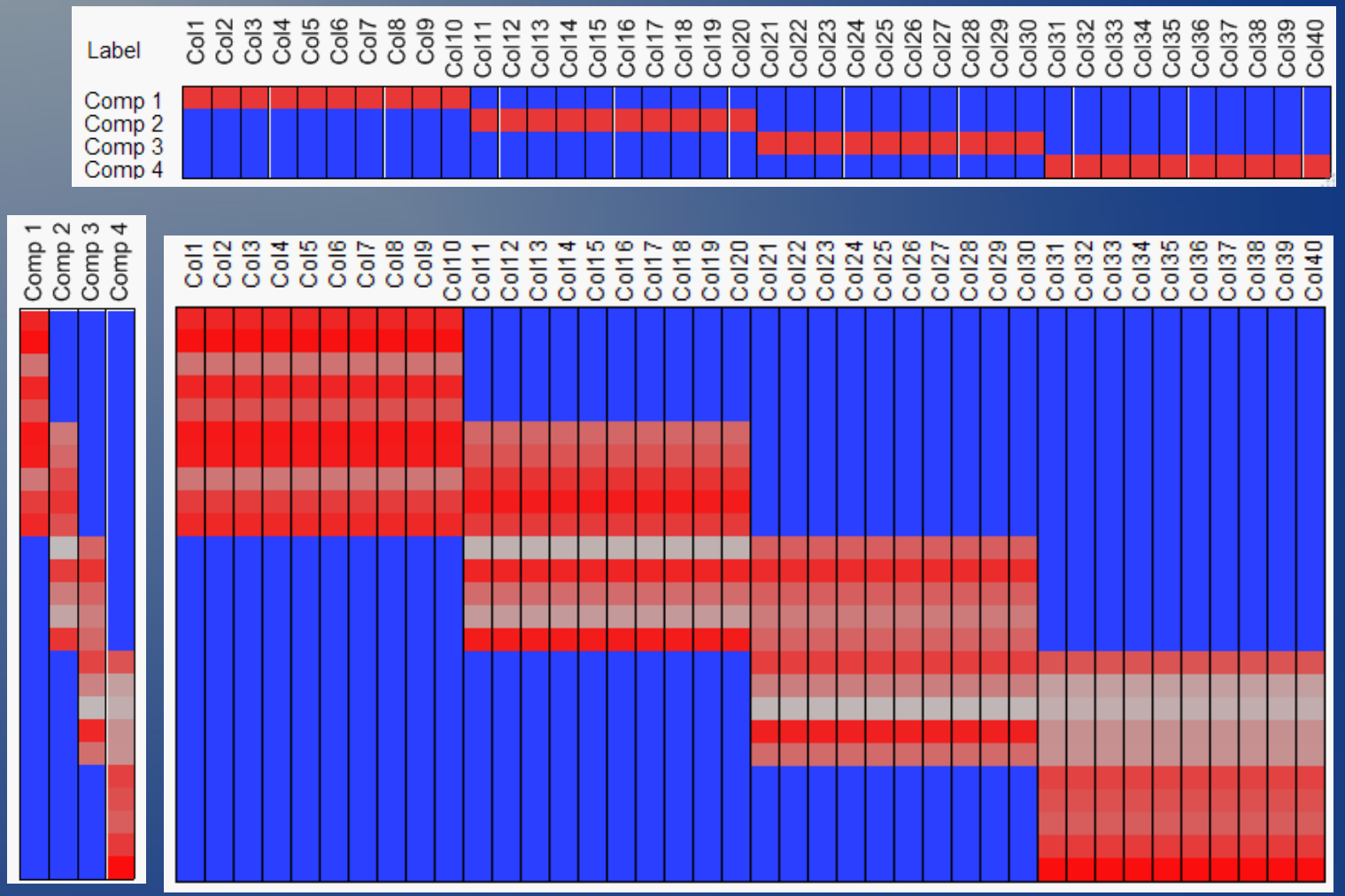
Crazy Synthetic Data Set



PCA – to predict



NMF – to explain



XXXX